Introduction to Biostatistics (BIOS:4120)
Breheny

Assignment 2
Due: Tuesday, February 6

1. True or False: In a hypothesis test, the null hypothesis can be summarized as "nothing is going on besides chance variation."

2. The null hypothesis is a hypothesis about (i) the sample (ii) the population.

3. Suppose that a scientist carries out 100 hypothesis tests. Unbeknownst to her, in all 100 cases, the null hypothesis is true. If she uses a cutoff of $p < 0.15$ to reject the null, about how many mistakes would you expect that she makes?

4. A drug has been developed that may reduce a person's cholesterol. Investigators are interested in estimating the amount by which the drug will reduce cholesterol, and in calculating a confidence interval. For each of the following, say whether the change will cause the confidence interval to get wider or get narrower:

   (a) The investigators decide to enroll more people in the study

   (b) More sophisticated lab techniques are used, allowing for more accurate measurement of cholesterol

   (c) The investigators want a 99% confidence interval instead of a 95% confidence interval

5. Which statistical procedure tells you more about the clinical significance of a study? (i) confidence intervals (ii) hypothesis tests

6. If one analyzes the clofibrate study as it was randomized, the p-value is 0.51.

   (a) True or false: There is a 51% probability that the null hypothesis is true.

   (b) True, false, or cannot be determined: A 75% confidence interval for the effectiveness of the drug would contain the null hypothesis value.

7. Read the short article "That Confounded p-value" [Lang1998].

   (a) The authors say that the information conveyed by p-values is "confounded." What do they mean by that?

   (b) On the second page of the article, the authors remark, "No one could infer the [confidence intervals] from the p-values. Given the [confidence intervals], no one needs these p-values." Would you agree with this sentiment? (There isn't necessarily a right or wrong answer here. You could probably make a valid argument either way; I just want to hear your opinion along with a reasonable justification)

   (c) The authors indicate a desire "to ban the reporting of all p-values from *Epidemiology*." Does this mean that they want to ban all statistical analysis?

8. In laboratory medicine and biology, it is now common to measure the expression levels of thousands of genes at once. So, for example, an investigator might collect samples from normal subjects and subjects with cancer in the hopes of finding genes that are significantly associated with cancer. The investigator is therefore testing a separate null hypothesis for each gene that is measured. Suppose that an investigator measures 2,000 genes, of which 20 are truly associated with cancer. Suppose further that the investigator's hypothesis tests have a Type I error rate of 5% and a Type II error rate of 20%.

   (a) Out of the 2,000 hypothesis tests that the investigator carries out, how many are type I errors?

   (b) How many are type II errors?

   (c) How many times did the investigator correctly reject the null hypothesis?

   (d) What was the investigator's false discovery rate?

   (e) If, for each gene, a 95% confidence interval was calculated for the association between the gene and cancer status, how many of those confidence intervals would contain the true association for that gene?

9. For each of the studies below, write a sentence summarizing the main findings (both are described in the "Study Design" lecture).

   (a) The Public Health Service's polio vaccine trial.

   (b) The Coronary Drug Project Research Group's clofibrate study.